Data Fusion for 3D Gestures Tracking using a Camera mounted on a Robot

Paulo Menezes†, Frédéric Lerasle‡, Jorge Dias†
†ISR-University of Coimbra, Polo II, 3030-290 Coimbra, Portugal
‡LAAS-CNRS, 17 av du colonel Roche, 31077 Toulouse cedex 4, France

Abstract

This article describes a multiple feature data fusion applied to an auxiliary particle filter for markerless tracking of 3D two-arm gestures by using a single camera mounted on a mobile robot. The human limbs are modelled by a set of linked degenerated quadrics which are truncated by pairs of planes also modelled as degenerated quadrics. The method relies on the projection of both the model's silhouette and local features located on the model surface, to validate the particles (associated configurations) which generate the best model-to-image fittings. Our cost metric combines robustly two imaging cues i.e. *model contours and colour or texture based patches located on the model surface, subject to 3D joint limits and also non self-intersection constraints. The results show the robustness and versatility of our data fusion based approach.*

1. Introduction

The development of personal robots is a motivating challenge in Robotics research. In this context, we have designed and implemented a mobile robot able to interact with its users. Gestures are especially valuable in crowded environments where speech recognition may be garbled or drowned out. They are natural and rich means that humans employ to communicate with each other. Mobile robot applications have to observe some high demanding requirements. First, on-board processing power is limited and care must be taken to design efficient algorithms. Second, the system must run at a speed which is comfortable for the human user. Third, as the robot may evolve in cluttered environments subjected to illumination changes, several hypotheses must be handled at each instant concerning the system parameters to be estimated, a robust integration of multiple visual cues is required to cope with the variations in both the environment and target appearance. We have implemented a markerless and monocular appearance-based gesture tracker that fulfills these requirements.

In the Vision community, many researchers have worked on markerless tracking systems [6] and on the problem of estimating the human pose from static images [4]. When just one camera is used, volumetric limbs' models [9, 2, 10, 4] are usually used to solve the ill-posed problem of estimating the 3D pose. The type of data used in matching between the models and the images varies from case to case, being edges [11, 2, 10], silhouette [2, 10], and motion [10], are the most used sources of information.

The computational weight is one of the limitations of most body trackers. Our two-arms gestures tracker is applied in a quasi real-time process. The method to handle model projection (see [5] for details), although being inspired from [11], is less time consuming. We focus, in this paper, on a new observation model that combines edges and motion cues, with local colour and texture patches on clothing or on the hands acting as natural markers.

Particle filtering is well-suited to our context as it makes no restrictive assumptions on the probability distributions and enables the easy fusion of diverse kinds of measurements.

Section 2 describes the likelihood function to be used on this Bayesian tracking. Tracking implementation and experiments on two-arm gestures are presented in sections 3 and 4. Section 5 summarizes our contribution and opens the discussion for future extensions.

2. Multiple cues fusion

Being this work based on the use of a particle filter as the base for the tracking mechanism, we explain in this section the construction of the measure functions to be used in the filter's weighting step. Knowing that each particle corresponds to an hypothesis of configuration for the 3D structure, the measurement step is responsible to evaluate how good is each of them.

Combining several cues may confer robustness w.r.t. temporary failures in some of the measurement processes, and enables the tracker to take advantage of the distinct features obtained from different information sources.

Given M measurement sources $(z_k^1, \ldots, z_k^M)$, the global measurement function can be factorized as

$p(z_k^1, \ldots, z_k^M|\mathbf{x}) \propto \prod_{m=1}^{M} p(z_k^m|\mathbf{x})$ This mixed weighting function is going to smooth some of the false peaks that may appear on each individual measure, and sharpens other ones. The result is that the tracker will be more robust as it will not be trapped by false peaks. The next subsections depict the measurements that are integrated in the particle weighting steps followed by some details regarding the implementation.

2.1. Image edges

A weighting factor can be computed for each particle after projecting the model contours and comparing them to the edges extracted from the image. The corresponding log-likelihood is classically computed using the sum of the squared distances between model points, uniformly placed on the model edges, and the nearest image edges [3]. In this implementation, the edge image is converted into a Distance Transform image, noted I_{DT}, which is used to peek the distance values. This has the advantage of both producing a smoother measure function and reducing the involved computations, when compared with the edge searching methods.

The edge-based likelihood given by $p(z_k^S|\mathbf{x}) \propto \exp\left(-\lambda_s \frac{D^2}{2\sigma_s^2}\right)$, $D = \sum_{j=0}^{N_p} I_{DT}(j)$ where j indexes the N_p model points uniformly distributed along each visible model projected segments, $I_{DT}(j)$ the associated value in the DT image, and λ_s a weighting factor. Figure 1.(a) plots this function for an example where the target is a 2D elliptical template corresponding coarsely to the head of the right subject in the input image. As it can be seen on this plot, for a cluttered background, the use of only shape cues for the model-to-image fitting is not sufficiently discriminant, as multiple peaks may appear.

2.2. Motion cues

In our context, the human limbs are expected to be moving, even if intermittently, in front of a background which is assumed to be static. Note that this assumption remains only valid if the camera is static during the process or undergoing a pure rotation where the background motion field can be estimated. To cope with cluttered scenes and reject false background attractors, we favour the moving edges, if they exist, as they are expected to correspond to the moving target. When the target is stopped, the static edges are not completely rejected, but only made less attractive than the moving ones. This is accomplished by using two DT images, noted I_{DT} and $I_{DT}^{'}$, where the new one is obtained by filtering out the static edges, based on the local the optical flow vector $\vec{f}(z)$. The new distance D is given by $D = \sum_{j=0}^{N_p} \min\left(I_{DT}(j), K.I_{DT}^{'}(j)\right)$ where K is a constant. Figure 1.(b) plots this more discriminant likelihood function for the example seen above. The target is still the right subject, who is assumed to be moving. The results show that the tracking is less disturbed by the background clutter, especially while the target is moving.

2.3. Local colour distributions

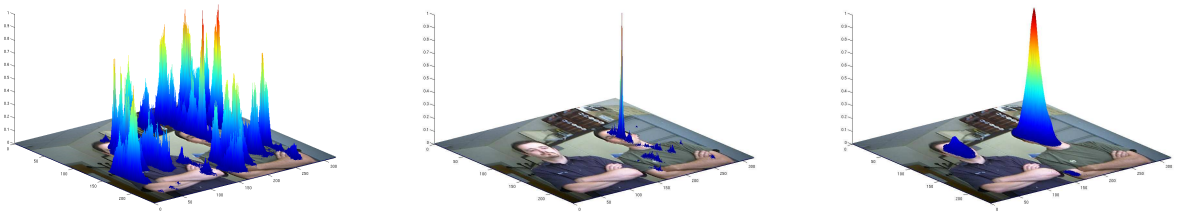
Clothes, normally, increase diversity of the patterns and of the colour sets that are found on one's body surface. They introduce contrasts between the colours of extremities, *e.g.* head, hands and feet, and the clothes that cover the trunk and arms. So, considering clothing patches of characteristic colour distributions, *i.e.* natural markers, seems very promising.

We denote the B-bin reference normalized histogram model in channel $c \in \{R, G, B\}$ by $h_{ref}^c = (h_{1,ref}^c, \ldots, h_{N_{bi},ref}^c)$. The colour distribution $h_x^c = (h_{1,x}^c, \ldots, h_{N_{bi},x}^c)$ of a region B_x corresponding to any state x is computed as $h_{j,x}^c = c_H \sum_{u \in B_x} \delta_j(b_u^c), j = 1, \ldots, N_{bi}$, where $b_u^c \in \{1, \ldots, N_{bi}\}$ denotes the histogram bin index associated with the intensity at pixel u in channel c of the colour image, δ_a terms the Kronecker delta function at a, and c_H is a normalisation factor. The colour likelihood model must be defined so as to favour candidate colour histograms h_x^c close to the reference histogram h_{ref}^c. The likelihood $p(z_k^C|\mathbf{x})$ is based on the Bhattacharyya coefficient [7] between the two histograms h_x^c and h_{ref}^c.

Considering several patches of distinct colours on the tracked limbs surface, the histogram-based modelling will capture them. We consider the partition $B_x = \bigcup_{p=1}^{N_R} B_{p,x}$ associated with the set of reference histograms $\{h_{p,ref}^c : c \in \{R, G, B\}, p = 1, \ldots, N_R\}$. By assuming conditional independence of the colour measurements, the multi-region colour likelihood becomes: $p(z^C|\mathbf{x}) \propto \exp(-\sum_c \sum_{p=1}^{N_R} \lambda_{p,c} \frac{D^2(h_{p,x}^c, h_{p,ref}^c)}{2\sigma_c^2})$ where the histogram $h_{p,x}$ is collected in the region $B_{p,x}$ and $\lambda_{p,c}$ the weighting factors. Figure 1.(c) plots this likelihood for the example seen above, where the target is a colour ROI corresponding to the head of the right subject.Being these measures quite discriminant in terms of colour/texture distributions, its use has shown to give very good results if discriminant colour spots on the body surface are tracked.

2.4. Stabilisation and collision detection

Despite the visual cues depicted above, ambiguities arise when certain model parameters cannot be inferred from the current image observations. For instance, when one arm is horizontal and the edge-base likelihood is used, rotation of the upper arm around its axial axis is unobservable, because



(a) (b) (c)

Figure 1. Likelihhoods regarding: (a) shape cue, (b) combined shape and motion, and (c) colour cue

the model projected contours remain static under this DOF. Instead of leaving these parameters unconstrained, and like in [10], we control these parameters with a stabiliser cost function that reaches its minimum on a predefined resting configuration $\mathbf{x}_{def}$. This enables the saving of computing efforts that would explore unobservable regions of the configuration space. In the absence of strong observations, the parameters are constrained to lie near their default values whereas strong observations unstick them from these default configurations. This can be expressed as a likelihood function for a state $\mathbf{x}$ as: $p_{st}(\mathbf{x}) \propto \exp(-\lambda_{st}||\mathbf{x}_{def} - \mathbf{x}||^2)$ This prior only depends on the structure parameters and the factor λ_{st} will be chosen in a way that the stabilising effect will be negligible for the whole configuration space with the exception of the regions where the other cost terms are constant.

Another point is that, as the estimation is based on a search on the configuration space, it would be desirable to a priori remove those regions that correspond to collisions between parts. Unfortunately it is in general not possible to define these forbidden regions in closed form so they could be rejected immediately during the sample phase. The result is that in the particle filter framework, it is possible that configurations proposed by some particles correspond to such impossible configurations, thus exploring regions in the configuration space that are of no interest. To avoid these situations, we use a binary cost function, that is not related to observations but only based on a collision detection mechanism. The corresponding likelihood function for a state $\mathbf{x}$ is $p_{coll}(\mathbf{x}) \propto \exp(-\lambda_{co} f_{co})$ with: $f_{co}(\mathbf{x}) = \{0, 1\}$ whether it corresponds to a collision or not.

3. Implementation

In its actual form, the system tracks eight degrees of freedom, *i.e.* four per arm. We assume therefore that the torso is coarsely fronto-parallel with respect to the camera while the position of the shoulders are deduced from the position of the face given by a dedicated tracker [1]. The patches are distributed on the surface model and their possible occlusions are managed during the tracking process. Our approach is different from the traditional marker-based ones because we do not use artificial but only natural colour or texture-based markers *e.g.* the two hands and ROIs on the clothes.

Regarding the particle filtering framework, we opt for the Auxiliary Particle Filter scheme introduced by Pitt and Shephard [8]. This allows to use some low cost measure or *a priori* knowledge to guide the particle placement, therefore concentrating them on the regions of interest of the state space. The associated measurement strategy is as follows: (1) particles are firstly located in good places of the configuration space according to rough correspondences between model patches and image features, and (2), on a second stage, particles' weights are fine-tuned using additional information from edges, motion, and colour patches.

4. Experiments and results

The above described approach has been implemented and evaluated over monocular images sequences acquired in various situations. Figures 2 and 3 show snapshots from two different sequences. The right sub-figures show the model projections superimposed to the original images for the mean state $E[\mathbf{x}_k^i]$ at frame k, while the left ones show its corresponding estimated configuration. The following examples combine measures that use the contours, three patches per arm, and the geometric constraints.

Due to the efficiency of the importance density and the relatively low dimensionality of the state-space, tracking results are achieved with a reasonably small number of particles *i.e.* $N_s = 400$ particles. In our unoptimised implementation, a PentiumIV-3GHz requires about $1s$ per frame to process the two arm tracking, most of the time being spent in observation function. To compare, classic systems take a few seconds per frame to process a single arm tracking.

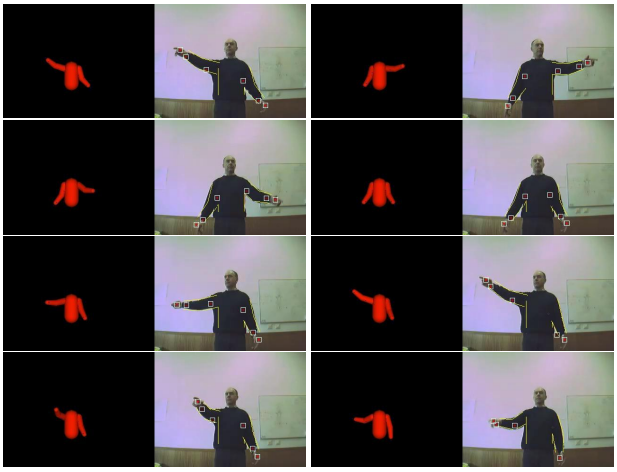


Figure 2. Tracking of pointing gestures

Figure 3. Tracking in the presence of clutter

5. Conclusion

We have presented a general Bayesian framework for multi-cue human limbs tracking using a single camera mounted on a mobile robot. After a first contribution, that was presented elsewhere [5] and, that deals with the method proposed to handle 3D model image projection and hidden removal efficiently, we propose a new model-image matching cost metric combining robustly visual cues and geometric constraints.

In our robotic context, no assumption about clothing appearance and environmental conditions can be made. Such variability is accounted by fusing in the global cost function and varying the degrees of confidence of the image cues. The estimation process is performed using the auxiliary particle filtering algorithm where the associated importance density is known to be the optimal strategy as it reduces at best the effects of degeneracy [8]. The integration of the measurement information in the sampling step enables to concentrate the filtering efforts where they really matter and so permits to reduce the number of particles.

Our experiments show, that the proposed framework is suitable for tracking 3D gestures and that the integration of multiple cues improve the tracker versatility. Our approach requires less computing power than most of existing ones, making possible its use in a quasi-real-time application.

References

[1] J.C. Barreto, P. Menezes, and J. Dias. Human robot interaction based on haar-like features and eigenfaces. In *ICRA'04*, 2004.

[2] J. Deutscher, A. Davison, and I. Reid. Automatic partitioning of high dimensional search spaces associated with articulated body motion capture. In *CVPR'01*, pages 669–676, 2001.

[3] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *ECCV'96*, pages 343–356, Cambridge, UK, April 1996.

[4] Mun Wai Lee and Isaac Cohen. Proposal maps driven MCMC for estimating human body pose in static images. In *CVPR'2004*, Washington, D.C., June 2004.

[5] P. Menezes, F. Lerasle, J. Dias, and R. Chatila. Single camera-based tracking of 3D gestures. In *ISR'05*, Tokyo, Japan, 2005.

[6] T. Moeslund and E. Granum. A survey on computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, 2001.

[7] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *IEEE (issue on State Estimation)*, 92(3), 2004.

[8] M.K. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446), 1999.

[9] H. Sidenbladh, M.J. Black, and D.J. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *ECCV'00*, pages 702–718, 2000.

[10] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *IEEE I. J. on Robotic Research*, 6(22):371–393, 2003.

[11] B. Stenger, P. R. S. Mendonça, and R. Cipolla. Model-based hand tracking using an unscented kalman filter. In *BMVC'01*, volume 1, pages 63–72, September 2001.